Recognition of synthesized vowel sequences in steady-state and sinusoidally amplitude-modulated noises

Yi Shen^{a)} and Dylan V. Pearson

Department of Speech and Hearing Sciences, Indiana University Bloomington, Bloomington, Indiana 47405, USA

(Received 28 July 2016; revised 21 February 2017; accepted 22 February 2017; published online 17 March 2017)

Modulation masking is known to impact speech intelligibility, but it is not clear whether the mechanism underlying this phenomenon is an invariant, bottom-up process, or if it is subjected to factors such as perceptual segregation and stimulus uncertainty thereby showing a top-down component. In the main experiment of the current study (Exp. II), listeners' ability to recognize sequences of synthesized vowels (i.e., the target) in sinusoidally amplitude-modulated noises (i.e., the masker) was evaluated. The target and masker were designed to be perceptually distinct to limit the top-down component of modulation masking. The duration of each vowel was either 25 or 100 ms, the rate at which the vowels were presented was either 1 or 6 Hz, and the masker modulation rate was varied between 0.5 and 16 Hz. The selective performance degradation when the target and masker modulation spectra overlap, as would be expected from modulation masking, was not observed. In addition, these results were able to be adequately captured using a model of energetic masking without any modulation processing stages and fitted only using the vowel-recognition performance in steady-state maskers, as obtained from Exp. I. Results suggest that speech modulation masking might not be mediated through an early-sensory mechanism. © 2017 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4978060>]

[GCS]

Pages: 1835–1843

I. INTRODUCTION

It has been repeatedly demonstrated that speech understanding in competing maskers is not solely determined by how much the target speech and masker overlap in terms of their excitations in the peripheral auditory system, i.e., energetic masking. In many situations, the similarity of the target and masker envelopes (i.e., overlap of the target and masker modulation spectra) can also lead to performance deficit, a phenomenon often referred to as modulation masking (e.g., Kwon and Turner, 2001). A number of recent studies have pointed out the important, and sometimes dominating, role of modulation masking for speech understanding in dynamically fluctuating maskers (Stone *et al.*, 2012; Stone and Moore, 2014; Fogerty *et al.*, 2016). For example, Fogerty *et al.* (2016) presented target sentences in noise maskers that were amplitude-modulated using the speech envelope. In addition, the masker was either time-compressed or time-expanded to systematically manipulate the modulation rate of the masker. It was found that a deficit in speech recognition occurred when the overall envelope fluctuation rates of the target and masker matched. This was in line with the idea that the auditory system exhibits a tuning characteristic in the modulation domain (e.g., Dau *et al.*, 1997; Yost *et al.*, 1989). However, this recent experimental evidence with regard to the role of modulation masking in the context of speech recognition is different from how modulation masking is traditionally conceptualized in psychoacoustic literature.

Traditionally, the concept of modulation masking has referred to the phenomenon that when a carrier sound is amplitude-modulated by two separate modulators, detection or discrimination based on one modulator could be undermined by the presence of the other (e.g., Houtgast, 1989). This phenomenon can be successfully captured by bottom-up models. For example, Dau *et al.* (1997) modeled modulation masking using a bank of bandpass modulation filters after the peripheral processing stages. In this model, modulation masking was treated similarly to energetic masking that occurs at the auditory periphery, and the shape of the modulation filters represents the tuning in the modulation domain.

In the context of speech recognition in noise, modulation masking often refers to a related but different phenomenon. That is, the recognition of target speech is more difficult when the competing masker has a fluctuating envelope, especially when the target and masker fluctuation rates are similar. For clarity, we will refer to this phenomenon as speech modulation masking in the following discussion. Because it is known that speech recognition involves top-down contributions, it may not be possible for speech modulation masking to be fully explained solely by a bottom-up mechanism. For example, it may be hypothesized that the top-down component of speech modulation masking may be originated from the perceptual fusion of the target and masker. When the target and masker exhibit similar envelope fluctuations (e.g., at similar fluctuation rates), the listener may be confused about the source identities and fail to selectively attend to the target speech.

According to this hypothesis, when the target and masker are perceptually distinct from each other, the

^{a)}Electronic mail: shen2@indiana.edu

top-down component of speech modulation masking would be reduced. In the current study, target and masker stimuli with highly regular and predictable temporal characteristics were employed to promote the subject's efficiency in selectively attending to the target. The target stimuli were sequences of synthesized steady-state vowels with a fixed rate (vowel rate), while the masker stimuli were sinusoidally amplitude-modulated noises. Under this special situation, if the effect of speech modulation masking, as observed by Fogerty *et al.* (2016), persists, then it is likely that bottom-up processes play a significant role in modulation masking for speech. On the other hand, if speech modulation masking becomes unmeasurable, then it supports the involvement of top-down processes in speech modulation masking.

In the current investigation, the masker modulation spectrum was determined by the masker modulation rate, while the target modulation spectrum was dependent on both vowel duration and vowel rate. These stimulus parameters (masker modulation rate, vowel duration, and vowel rate) were manipulated to systematically vary the degree of overlap between the modulation spectra of the target and masker. However, it is expected that these stimulus parameters could potentially affect vowel recognition without the involvement of speech modulation masking. First, as pointed out by Miller and Licklider (1950), at low masker modulation rates, the duration of individual masker envelope peaks could be sufficiently long to encompass the entire words. For the vowel sequence stimuli used in the current study, the occurrence of whole-word masking would be more frequent when the duration of each vowel becomes shorter than the period of the masker modulation. Therefore, the dependency of vowel-recognition on masker modulation rate is, in part, determined by the relationship between the vowel duration and the masker modulation rate. Second, it is expected that recognizing more rapid vowel sequences will be more cognitively demanding. For example, it is known that when speech is time-compressed into a fast rate, it could undermine its intelligibility, especially for older listeners (e.g., Stine *et al.*, 1986; Gordon-Salant and Fitzgibbons, 1993; Letowski and Poch, 1995; Vaughan and Letowski, 1997). For vowel sequences presented in quiet, previous studies have suggested that the listeners' ability to identify temporal order of vowels in a sequence is degraded for fast vowel rates (Warren and Warren, 1970; Thomas *et al.*, 1971; Dorman *et al.*, 1975).

To control for these potential confounds, the involvement of speech modulation masking was determined by comparing vowel-recognition performance in amplitude-modulated maskers measured experimentally against the predictions from a model of energetic masking that did not include any modulation processing stages. The model was implemented here to account for the potential effects of masker modulation rate, vowel duration, and vowel rate on speech intelligibility that are not related to speech modulation masking. The model was based on the hypothesis that listeners are able to take advantage of the momentary high signal-to-noise ratio (SNR) segments during the masker envelope dips, allowing the listeners to glimpse the target speech (Buus, 1985; Howard-Jones and Rosen, 1993; Rhebergen and Versfeld, 2005). In the model, listeners' ability to "listen in the dips" is limited

by the sluggishness of the auditory system at high masker modulation rates (e.g., Dubno *et al.*, 2002, 2003; Rhebergen and Versfeld, 2005; Jin and Nelson, 2006; George *et al.*, 2006; Gregan *et al.*, 2013). If a deviation between the experimental and modeling data occurred when the target and masker modulation spectra overlap, then it would be considered as evidence for speech modulation masking.

In summary, the current study investigated whether speech modulation masking significantly contributes to speech intelligibility even when the potential top-down component of speech modulation masking is limited by using perceptually distinct targets and maskers. If speech modulation masking plays a significant role in speech intelligibility, then as the modulation spectra of the target and masker are systematically manipulated the performance should vary according to the degree of overlap between the two modulation spectra. This signature of speech modulation masking was assessed by comparing experimental data to a model of energetic masking, which was used to control for the effects of stimulus manipulations that were unrelated to speech modulation masking. Following, the results from a control experiment (Exp. I) and the main experiment (Exp. II) will be presented first, which measured the recognition of vowel sequences in steady-state and amplitude-modulated maskers, respectively. Following the presentation of the experimental data, a model of energetic masking will be described and comparisons between the experimental data (from Exp. II) and model predictions (with the model fitted using Exp. I data) will be conducted. It will be shown that the involvement of speech modulation masking was not observed in the current study, which provided evidence for the involvement of top-down processes in speech modulation masking.

II. EXPERIMENT I: VOWEL SEQUENCES IN STEADY-STATE NOISES

A. Methods

1. Listeners

Six listeners (four females), between the ages of 18 and 24 yrs, were recruited from the student population at Indiana University. All listeners had audiometric thresholds equal to or better than 15 dB HL between 250 and 4000 Hz in both ears. For each listener, the ear with a lower pure tone average (PTA) threshold (mean of the hearing thresholds at 0.5, 1, and 2 kHz) was tested. If the PTA thresholds were the same for both ears, the left ear was tested. All of the listeners had practiced for more than 4 h on an experiment similar to the present experiment. Three of the listeners had previously participated in other psychoacoustic experiments.

2. Stimuli

Listeners identified sequences of five synthesized vowels in steady-state noise while the vowel duration and vowel rate (in terms of vowels per second) were systematically varied [see Figs. 1(A) and 1(B) for schematics of vowel sequences in steady-state noise for low and high vowel rate, respectively]. Each vowel in the sequence was drawn, with replacement, from a list of five possible synthesized vowels

FIG. 1. The schematic of the stimuli for a (A) low- and (B) high-rate vowel sequence with steady state masker (used in Exp. I) as well as a (C) high-rate vowel sequence with amplitude-modulated masker (used in Exp. II). Panel (D) illustrates the modulation spectra (solid curves) of the target vowel sequences for the 1-Hz (left) and 6-Hz (right) vowel rates. The masker modulation rates tested in Exp. II are indicated using circles.

/a, i, u, æ, ʊ/ with equal probability. The synthesized vowels were constructed using a Klatt synthesizer (Klatt, 1980) with a fixed fundamental frequency of 125 Hz, and each synthesized vowel contained five formant peaks. The bandwidths of the five formants were 90, 110, 170, 250, and 300, from the first to the fifth formant. The vowel identities were determined by formant frequencies. The formant frequencies for the five possible vowels, shown in Table I, were set according to Assmann and Summerfield (1994) and Lentz and Marsh (2006). The presentation level for each vowel was 65 dB SPL. In separate conditions, the vowel durations were 12.5, 25, 50, and 100 ms, and the vowel rates were 1 and 6 Hz. Each vowel was gated on and off using 1-ms raised-cosine ramps. In all conditions, the vowel sequence was presented with a simultaneous masker. The masker was a broadband noise filtered to match the long-term spectrum of the vowel sequence. The masker was gated on 500 ms before the onset of the vowel sequence and gated off 500 ms following the offset of the vowel sequence using 1-ms raised-cosine ramps. The masker level on each trial was set according to the SNR for that trial.

The stimuli were generated digitally and presented at a sampling rate of 22.05 kHz using custom software running in MATLAB (The MathWorks, Inc., Natick, MA), which was also used for experimental control and the collection of responses. The stimuli were presented monaurally in the test ear through a 24-bit sound card (MOTU Microbook II, Mark of the Unicorn, Inc., Cambridge, MA) and headphones (HD-280 Pro, Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany). During the experiment, listeners were seated in a single-walled sound-attenuating booth (Eckel AB-4250, Eckel Industries Inc., Cambridge, MA).

3. Procedure

Before data collection, each listener ran through a practice protocol in which the vowel rates were 1 and 6 Hz and the vowel duration was fixed at 100 ms. Half of the listeners started with the 1-Hz vowel rate first, while the other half

started with the 6-Hz vowel rate. Three experimental blocks were run for each of the vowel rates. For each trial the listener reported the identity of the presented vowels in order by selecting them from a list of the five vowels used in the experiment. Responses were given immediately after the sequence presentation in each trial. After each trial the listener received visual feedback showing the proportion of vowels correctly identified for that trial. These experimental blocks were constructed using the same procedure as the main experiment, which will be described below.

When the data collection began, each listener was tested with the two vowel rates in the same order as in the practice protocol. For each vowel rate, the four vowel durations were tested in random order and after one experimental block had been run for each vowel duration, the process was repeated two more times with independent random draws. Within each experiment block, vowel recognition was tested using the method of constant stimuli. Signal-to-noise ratios of -10 , -7.5 , -5 , -2.5 , 0 , 2.5 , and 5 dB were tested in random order, and the measurements at these SNRs were repeated six times with independent random sequences each time. At the completion of the experiment, all experimental trials under the same combination of vowel duration and vowel rate were pooled together (126 trials total), which contributed to the estimation of the psychometric functions relating the percent correct in vowel recognition to SNR.

B. Results

Figure 2 plots the psychometric functions for vowel recognition from Exp. I. Results for individual listeners are arranged in rows, while the two vowel rates are arranged into the two columns. Within each panel, the four vowel durations are indicated using different symbols. The average psychometric functions are shown in Fig. 3. In both figures, the vowel-recognition performance is indicated in rationalized arcsine units or RAU (Studebaker, 1985). In general, as the SNR increased, the performance in vowel recognition improved. This performance improvement was more evident for the lower vowel rate and for the longer vowel durations. For example, with the 1-Hz vowel rate and 100-ms vowel duration (diamonds in the left panels in Fig. 2), the performance approached 100 RAU for an SNR of 5 dB. For the same vowel duration with the 6-Hz vowel rate (diamonds in the right panels in Fig. 2), the performance did not exceed beyond 83 RAU for the range of SNRs tested. To quantify the shape of the psychometric function, the data for each of

TABLE I. Formant frequencies (in Hz) for the vowel stimuli.

Vowel	/a/	/i/	/æ/	/ɜ:/	/u/
F1	750	250	750	450	250
F2	1050	2250	1450	1150	850
F3	2950	3050	2450	1250	2250
F4	3350	3350	3350	3350	3350
F5	3850	3850	3850	3850	3850

FIG. 2. The psychometric function relating percent correct in vowel recognition (in RAU) to SNR from Exp. I for individual listeners (in separate rows). Data for the 1- and 6-Hz vowel rates are arranged into the left and right columns, respectively. The four vowel durations (12.5, 25, 50, and 100 ms) are indicated using different symbols. The horizontal dashed line in each panel indicates the expected chance performance.

the eight combinations of vowel rate and duration were fitted with a logistic function using a maximum-likelihood algorithm (e.g., [Wichmann and Hill, 2001](#)). The fitted function took the form

$$\text{Percent Correct} = \gamma + (100 - \gamma - \lambda) [1 + e^{-\beta(\text{SNR} - \alpha)}]^{-1}. \quad (1)$$

According to Eq. (1), the psychometric function is a sigmoidal function relating the SNR (x axis, with a unit of dB) to the percent correct (y axis, with a unit of %). Along the y axis, the function ranges from γ to $1 - \lambda$, where γ indicates the chance performance (i.e., 20%), and λ indicates the distance between the upper performance plateau and 100% correct. Note that the parameter λ is needed here to capture the possibility that the vowel-recognition performance may never reach 100% even at high SNRs. Along the x axis, the center of the function is located at α , therefore, a higher α corresponds to a higher performance threshold. We will refer to α as the threshold parameter. The slope of the

FIG. 3. The psychometric function relating percent correct in vowel recognition (in RAU) to SNR from Exp. I averaged across six listeners. Data for the 1- and 6-Hz vowel rates are arranged into the left and right columns, respectively. The four vowel durations (12.5, 25, 50, and 100 ms) are indicated using different symbols. The horizontal dashed line in each panel indicates the expected chance performance. Error bars indicate $\pm$ one standard error of the mean.

psychometric function is determined by parameter β , and a greater value of β indicates a steeper slope of the psychometric function. We will refer to β as the slope parameter.

For the threshold parameter α , a repeated-measures analysis of variance (ANOVA) indicated a significant effect of vowel duration [$F(1.05, 5.24) = 7.513, p = 0.038$, Greenhouse-Geisser corrected], while the effect of vowel rate was not found to be significant [$F(1, 5) = 7.917, p = 0.702$] neither was the interaction between vowel rate and vowel duration [$F(1.02, 5.11) = 1.042, p = 0.403$, Greenhouse-Geisser corrected]. *Post hoc* pairwise comparisons showed that the α estimate for the 50-ms vowel duration was significantly lower than the estimate for 12.5 ms, and the α estimate for the 100-ms vowel duration was significantly lower than the estimate for 50 ms. For the slope parameter β , a repeated-measures ANOVA indicated no significant effect of vowel rate [$F(1, 5) = 0.245, p = 0.641$] or vowel duration [$F(1.36, 6.82) = 0.368, p = 0.628$, Greenhouse-Geisser corrected]. There was no significant interaction between vowel rate and vowel duration [$F(1.23, 6.12) = 4.535, p = 0.072$, Greenhouse-Geisser corrected]. For λ , a repeated-measures ANOVA indicated a significant effect of vowel rate [$F(1, 5) = 23.843, p = 0.005$], while the effect of vowel duration was not found to be significant [$F(1.18, 5.90) = 0.100, p = 0.802$, Greenhouse-Geisser corrected] neither was the interaction between vowel rate and vowel duration [$F(3, 15) = 0.483, p = 0.699$].

The effects of vowel duration and vowel rate demonstrated in the current experiment were observed in the absence of any imposed amplitude modulation on the masker. The data collected in Exp. I served as a baseline and were used, along with a model of energetic masking (see Sec. IV for details), to identify the potential involvements of speech modulation masking in the main experiment (i.e., Exp. II).

III. EXPERIMENT II: VOWEL SEQUENCES IN SINUSOIDALLY AMPLITUDE-MODULATED NOISES

A. Methods

Four listeners (three females) were included in the current experiment; all had also participated in Exp. I

previously. The stimuli for Exp. II were identical to those of Exp. I, unless stated otherwise. The vowel durations were 25 and 100 ms, and the vowel rates were 1 and 6 Hz. In all conditions, the simultaneous masker was sinusoidally amplitude-modulated at a modulation depth of 100% [see Fig. 1(C) for a schematic of a vowel sequence embedded in a sinusoidally amplitude-modulated noise]. The initial phase of the modulator was randomly drawn from a uniform distribution spanning 0 through 2π . In separate conditions, the masker modulation rates were 0.5, 1, 2, 4, 6, 8, and 16 Hz. Figure 1(D) illustrates the relationship between the masker modulation rates (indicated by circles) and the modulation spectra (solid curves) of the target vowel sequences for the 1-Hz (left) and 6-Hz (right) vowel rates. Note that the two vowel rates give rise to distinct modulation spectra, and it is expected that if speech modulation masking dominates vowel recognition in the current experiment, the function relating vowel recognition performance to masker modulation rate should also be different depending on vowel rate. There were a total of 28 experimental conditions (two vowel durations $\times$ two vowel rates $\times$ seven masker modulation rates) in Exp. II, and these conditions were tested using the same methods as Exp. I in the following sequence. Each listener began with the vowel rate that was tested first in Exp. I. For this vowel rate, the seven masker modulation rates were tested in random order. Within each masker modulation rate, the two vowel durations were tested also in random order using two experimental blocks. Each experimental block tested vowel recognition at seven SNRs in random order and repeated six times using the identical procedures as in Exp. I. Once one experimental block was run for each combination of masker modulation rates and vowel durations, the above stated process was repeated two more times with independent random sequences. After the three repetitions were complete, the second vowel rate was tested using the same randomized block design.

B. Results

The psychometric functions relating to vowel-recognition performance (in RAU) to SNR in all 28 conditions were fairly consistent across the four listeners. The standard deviation was about 9.2 RAU on average. Figure 4 plots the average performance as functions of SNR and masker modulation rate for vowel durations of 25 (left panels) and 100 (right panels) ms and for vowel rates of 1 (top panels) and 6 (bottom panels) Hz. The overall performance was poorer for the 6-Hz vowel rate than the 1-Hz vowel rate. Overall, performance dropped with decreased SNR, but the decrease was larger in the 1-Hz condition. For the 25-ms vowel duration, the psychometric function did not seem to depend on masker modulation rate. For the 100-ms vowel duration, on the other hand, performance improved as the masker modulation rate increased for low SNRs but not for high SNRs. At the highest SNRs, the performance was not affected by the masker amplitude modulation and was mainly determined by the vowel rate.

A repeated-measures ANOVA was conducted for each combination of the vowel rate and vowel duration to confirm

the observed effects of SNRs and masker modulation rate on vowel-recognition performance (in RAU). For the 25-ms vowel duration, the only significant main effects were from SNR [$F(6, 18) = 121.283$, $p < 0.001$ for the 1-Hz vowel rate; $F(6, 18) = 58.641$, $p < 0.001$ for the 6-Hz vowel rate], while there was no significant effect of masker modulation rate [$F(6, 18) = 2.351$, $p = 0.075$ for the 1-Hz vowel rate; $F(6, 18) = 0.803$, $p = 0.581$ for the 6-Hz vowel rate] nor were there significant interactions between SNR and masker modulation rate [$F(36, 108) = 0.913$, $p = 0.612$ for the 1-Hz vowel rate; $F(36, 108) = 0.765$, $p = 0.820$ for the 6-Hz vowel rate]. For the 100-ms vowel duration, the main effects of SNR were significant [$F(6, 18) = 58.079$, $p < 0.001$ for the 1-Hz vowel rate; $F(6, 18) = 30.897$, $p < 0.001$ for the 6-Hz vowel rate], as were the effects of masker modulation rate [$F(6, 18) = 29.263$, $p < 0.001$ for the 1-Hz vowel rate; $F(6, 18) = 18.252$, $p < 0.001$ for the 6-Hz vowel rate]. There were also significant interactions between SNR and masker modulation rate [$F(36, 108) = 5.105$, $p < 0.001$ for the 1-Hz vowel rate; $F(36, 108) = 4.669$, $p < 0.001$ for the 6-Hz vowel rate].

Therefore, it seems that the benefit of higher masker modulation rates compared to lower rates, reflected as the significant effects of masker modulation rate, only occurred when the vowel duration was sufficiently long. This is potentially because the 25-ms vowel duration was always shorter than the period of the masker amplitude modulation (62.5–2000 ms), leading to a higher probability ($\sim 50\%$) of the whole vowel being within a single masker envelope peak. At low SNRs, this whole-word masking effect would cause the performance to approach 50% correct. On the other hand, for the 100-ms vowel duration whole-word masking would only occur at low masker modulation rates.

For both the 1- and 6-Hz vowel rates, the performance pattern across masker modulation rates did not exhibit evident resemblance to the target modulation spectra [Fig. 1(D)]. For example, selective performance degradations near the 6-Hz masker modulation rate were not observed for the 6-Hz vowel rate. This could mean that the contributions to the recognition of vowel sequences from the speech modulation masking mechanisms might be limited in the current experiment. This will be discussed in greater detail with the help of an energetic masking model in Sec. IV.

IV. PREDICTIONS FROM A MODEL OF ENERGETIC MASKING

To further investigate the potential involvement of speech modulation masking in Exp. II, a model of energetic masking was implemented and fitted using data collected from Exp. I. As this model contained no processing stages that related to speech modulation masking, if a close agreement between model predictions and the data of Exp. II could be found, it would mean that the amount of speech modulation masking was very limited.

A common approach to model energetic masking during speech recognition in background noise was based on the evaluation of short-term SNRs (e.g., [Rhebergen and Versfeld, 2005](#); [Rhebergen et al., 2006](#); [Bernstein and Grant, 2009](#); [Ma et al., 2009](#); [Shen et al., 2015](#)). This approach,

FIG. 4. (Color online) The psychometric function relating percent correct in vowel recognition (in RAU) to SNR and masker modulation rate from Exp. II averaged across six listeners. Percent correct is plotted along the z axis while SNR is plotted along the x axis and the modulation masker rate is plotted along the y axis. The vowel rate is split across the top (1 Hz) and bottom (6 Hz) panels and vowel duration is split across the left (25 ms) and right (100 ms) panels. A subset of these data are also replotted in the left panels of Fig. 6.

based on the hypothesis that listeners are able to take advantage of the momentary high-SNR periods during the dips of the masker envelope to glimpse speech segments, was implemented through a simple model of the dip-listening hypothesis following [Shen et al. \(2015\)](#). Compared to the dip-listening models proposed previously (e.g., [Rhebergen and Versfeld, 2005](#)), this model computes the short-term SNRs analytically rather than numerically, taking advantage of the fact that the target and masker had deterministic temporal envelopes in the current experiment. For maskers with sinusoidal amplitude modulation and speech segments that are temporally steady, the local, short-term SNR is related to the nominal SNR

$$\text{SNR}_{\text{local}}(t, \varphi) = \text{SNR}_{\text{nominal}} - 10 \log \left\{ \frac{2}{3\Delta T} \int_t^{t+\Delta T} [1 + \sin(2\pi f_m t + \varphi)]^2 dt \right\}, \quad (2)$$

where t indicates time relative to the onset of the speech segment, φ is the initial phase of the masker modulator, ΔT is the duration of the temporal integration window, and f_m is the masker modulation rate. Figure 5 shows how $\text{SNR}_{\text{local}}$ (bottom row) is related to the target (top row) and masker (middle row) envelopes. The left and right panels are two illustrations associated with low and high masker modulation rates, respectively. Because the $\text{SNR}_{\text{local}}$ is expected to be $-\infty$ dB in the absence of the target, it is only evaluated within the vowel duration T (labeled in the top left panel of Fig. 5), i.e., $0 < t < T$.

If the psychometric function relating the nominal SNR to the percent correct recognition for steady-state maskers, i.e., $\text{PF}_{\text{UM}}(\text{SNR}_{\text{nominal}})$, is available (e.g., from fitting the psychometric function to the Exp. I data), then the expected psychometric function for sinusoidally amplitude-modulated maskers can be written as

$$\text{PF}_{\text{MOD}}(\text{SNR}_{\text{nominal}}) = \frac{1}{2\pi T'} \int_0^{2\pi} \int_{T'} \text{PF}_{\text{UM}}[\text{SNR}'_{\text{local}}(t, \varphi)] \times dt d\varphi, \quad (3)$$

where $\text{SNR}'_{\text{local}}$ is the value from Eq. (2) but bounded to be between -15 and 15 dB, according to the ANSI standard for the computation of the Speech Intelligibility Index ([ANSI, 2007](#)), and T' is the duration of the speech segment that contributes to the recognition of the word. The value of T' is between 0 and T . Smaller T' values indicate shorter glimpses, while T' values close to T indicate long glimpses.

Let us consider T' as a free parameter of the model, for a given value of T' , a criterion can be set on $\text{SNR}_{\text{local}}$ (see the horizontal dashed lines in the bottom panels of Fig. 5). This criterion is set so that the total duration during which the

FIG. 5. The derivation of the local SNR ($\text{SNR}_{\text{local}}$, bottom panels) from the envelopes of the target (top panels) and masker (middle panels) in the model of energetic masking used in the current study. Illustrations corresponding to low and high masker modulation rates are shown in the left and right panels, respectively. The vowel duration T is labeled in the top left panel. In each of the bottom panels, the horizontal dashed line indicates a criterion that defines the periods of T' within the vowel duration with the highest $\text{SNR}_{\text{local}}$ (shaded region). In the model, $\text{SNR}_{\text{local}}$ within this period contributes to the prediction of vowel-recognition performance.

$\text{SNR}_{\text{local}}$ is above the criterion is T' (the shaded durations in the bottom panels of Fig. 5). The evaluation of the time-domain integral in Eq. (3) is based on these durations. Equation (3) represents expected performance taking into account the temporally variable $\text{SNR}_{\text{local}}$ and the variable phase relationship between the target and masker envelopes (represented in φ).

The above described model was implemented to investigate whether such a simple dip-listening model could explain the data obtained from Exp. II, using only the psychometric function estimated from Exp. I. For each listener who participated in both experiments, the psychometric function for unmodulated maskers PF_{UM} was estimated by fitting a logistic function to all experimental trials collected in Exp. I using a maximum-likelihood algorithm. The logistic psychometric function was in the form given in Eq. (1), except that

$$\alpha = \alpha_0 + \alpha_1 \log T, \quad (4)$$

and

$$\lambda = \begin{cases} \lambda_1 & \text{if } R = 1 \\ \lambda_6 & \text{if } R = 6, \end{cases} \quad (5)$$

where T indicates the vowel duration in seconds, R indicates the vowel rate in Hz, and the psychometric function could be fully described by the free model parameters $\phi = \{\alpha_0, \alpha_1, \beta, \lambda_1, \lambda_6\}$ and the stimulus parameters $\text{SNR}_{\text{nominal}}, T$, and R , i.e., $\text{Percent Correct} = \text{PF}_{\text{UM}}(\phi, \text{SNR}_{\text{nominal}}, T, R)$. The specific formulation of the psychometric function given in Eqs. (4) and (5) was based on the significant effect of vowel duration on α and significant effect of vowel rate on λ found in Exp. I (see Sec. II B). The fitted parameters are listed in Table II.

To predict performance for Exp. II, the estimated psychometric function was inserted into Eq. (3), forming the predicted psychometric function for amplitude-modulated maskers. Figure 6 plots the measured (left) and predicted (right) percent correct vowel recognition for a subset of SNRs tested in Exp. II ($-10, -5, 0$, and 5 dB). The experimental data are the average across the four listeners, while the modeling data are based on the average psychometric-function parameters estimated from Exp. I (i.e., the bottom row in Table II). The four combinations of vowel rate and vowel duration were arranged in rows. For the modeling results shown in the figure, a sample set of parameter values

TABLE II. The psychometric function parameters [Eqs. (1), (4), and (5)] estimated from the four listeners who participated in both Exps. I and II. The average parameters across the listeners are listed in the bottom row.

	α_0	α_1	β	λ_1	λ_6
S1	-5.15	-19.70	0.47	7.22	33.14
S2	-4.05	-16.41	0.57	1.05	24.70
S4	-3.91	-15.17	0.58	0.94	27.67
S5	-4.00	-16.99	0.49	23.93	40.76
AVG	-4.28	-17.07	0.53	8.29	31.57

FIG. 6. Left: The mean percent correct in identifying each vowel across listeners as functions of the masker modulation rate from Exp. II. Only a subset of the SNRs (different symbols) is shown for visual clarity. Various combinations of vowel duration and vowel rate are arranged in rows. Right: Same as the left panels but for model predicted results. These model predictions were generated using a simple dip-listening model based on the assessment of short-term SNRs. Data from Exp. I were used to fit the model.

was chosen with the duration of the temporal integration window set to $\Delta T = 8$ ms, and T' set to 25 ms. The experimental and modeling data showed close resemblance, suggesting that a short-term-SNR based model (i.e., a model of energetic masking) could be used to explain the relationship between vowel-sequence recognition in steady-state (Exp. I) and amplitude-modulated (Exp. II) noise maskers.

The model implemented here had two free parameters ΔT and T' . The model parameter ΔT only affected results to a limited degree and it would mainly impact the model prediction at masker modulation rates higher than those tested in Exp. II [see Shen *et al.* (2015) for some demonstrations of the effects of ΔT]. On the other hand, the parameter T' influenced the model prediction in a systematic manner. As T' increased, the predicted percent correct at low SNRs and high masker modulation rates decreased. This could be understood as the following: The current model prediction was based on segments of the vowel duration with the highest SNRs (see the shaded durations in the bottom panel of Fig. 5), for smaller T' values, these segments were associated with the masker envelope dips. On the other hand, as the T' value approached the

FIG. 7. The effect of the parameter T' on the RMS deviation between the experimental and model data for Exp. II. Results are shown for separate combinations of vowel rate and vowel duration (symbols in the left panel), as well as for the pooled data across experimental conditions (solid curve in the right panel). The value of T' that provided the smallest amount of overall RMS deviation is indicated using a dashed vertical line in the right panel.

vowel duration, vowel segments during the envelope peaks of the masker began to contribute to the predicted performance, reducing the expected performance.

To find the value of T' that provided the closest prediction of the vowel-recognition performance in Exp. II, the root-mean-square (RMS) deviation from the experimental data to the model predictions was computed as a function of T' . As in Fig. 6, the model predictions were based on the average psychometric-function parameters from Exp. I, and the experimental data were based on the average percent correct across listeners. When T' was greater than 25 ms, it was capped at 25 ms for the 25-ms vowel duration. The left panel of Fig. 7 plots the RMS deviation for each of the four combinations of vowel rate and vowel duration, while the right panel plots the overall RMS deviation. The RMS deviation was less than 6% across T' values from 10 to 50 ms for all conditions except for the 6-Hz vowel rate and 100-ms vowel duration where the RMS deviation was between 9% and 10%. Collapsing the four vowel rate and vowel duration conditions together, the T' value that led to the minimum overall RMS deviation was 32 ms (vertical dashed line in the right panel of Fig. 7). This value agreed with the duration required for identifying vowels in isolation (Thomas *et al.*, 1970).

V. GENERAL DISCUSSION

The model of energetic masking utilized here did not include any mechanisms that represented speech modulation masking. The close agreement between the data and model predictions suggested that speech modulation masking might not always play a dominant role in speech intelligibility in temporally fluctuating maskers, specifically when the target and masker have highly regular and predictable temporal characteristics that improves the subject's ability to attend to the target. The lack of evident effects of speech modulation masking in the current study could mean that the speech modulation masking effects observed from other studies of the speech-in-noise perception (e.g., Stone *et al.*, 2012; Stone and Moore, 2014; Oxenham and Kreft, 2014; Fogerty *et al.*, 2016) might not be governed by a stimulus-driven

mechanism, hard-wired into the early auditory processing stages. Rather than being invariant, the amount of speech modulation masking is sensitive to the conditions and methods with which it is measured.

In the current study, the vowel sequences were perceptually distinct from the masking noise, because (1) they carried a strong periodicity cue in their temporal fine structures (e.g., Lorenzi *et al.*, 2006; Hopkins and Moore, 2009; Bernstein and Brungart, 2011), and (2) their temporal envelopes were highly predictable (e.g., Jones *et al.*, 1981; Bendixen *et al.*, 2010). Well-segregated target and masker could facilitate listeners to adopt optimal glimpsing strategies, i.e., glimpsing only during the masker envelope dips. Therefore, one might hypothesize based on the current modeling results that, for well-segregated target and maskers, the intelligibility of the target speech in temporally fluctuating maskers could be predicted from the intelligibility in steady-state maskers using models based on the assessment of short-term SNRs.

The experimental paradigm developed in the current study, i.e., the recognition of synthesized vowel sequences under simultaneous masking, may provide a useful tool for future studies. Since the energetic masking effect for well-segregated targets and maskers can be quantitatively predicted under this paradigm, central processes involved in speech recognition may be investigated from the deviation of the recognition results from model predictions. For example, one may systematically remove the available acoustic cues for the perceptual segregation of the target and masker (such as temporal fine structure) and investigate whether an additional amount of masking, beyond energetic masking, would be observed. These future investigations would contribute to our understanding of speech intelligibility under complex listening environments.

VI. SUMMARY

Listeners' capabilities in recognizing synthesized vowels in amplitude-modulated noise maskers were investigated with masker modulation rate, vowel rate, and vowel duration systematically varied (in Exp. II). Although the manipulations of these stimulus parameters differentially changed how much the modulation spectra of the target and masker overlap, the pattern of results expected from speech modulation masking was not observed in the current study. The effects of vowel rate and vowel duration on vowel-recognition performance closely followed those observed when the vowel sequences were presented in steady-state maskers (in Exp. I). The effect of the masker modulation rate in Exp. II was dominated by whole-word masking. Therefore, for perceptually distinct target and masker with high temporal regularities, the influences of speech modulation masking to speech recognition may be quite limited. Speech modulation masking's limited impact for these target masker pairs leads us to believe that it may not originate from a single invariant, bottom-up mechanism.

ACKNOWLEDGMENTS

This work was supported by the NIH Grant No. R21 DC013406 (MPIs: V. M. Richards and Y. Shen). Christine

Clements provided assistance for data collection. Elsie B. Hayslip and Allison Kern contributed to the data analysis and the preparation of the manuscript.

- ANSI (2007). ANSI S3.5-2007, *American National Standard Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).
- Assmann, P. F., and Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels," *J. Acoust. Soc. Am.* **95**, 471–484.
- Bendixen, A., Denham, S. L., Gyimesi, K., and Winkler, I. (2010). "Regular patterns stabilize auditory streams," *J. Acoust. Soc. Am.* **128**, 3658–3666.
- Bernstein, J. G., and Brungart, D. S. (2011). "Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio," *J. Acoust. Soc. Am.* **130**, 473–488.
- Bernstein, J. G., and Grant, K. W. (2009). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **125**, 3358–3372.
- Buus, S. (1985). "Release from masking caused by envelope fluctuations," *J. Acoust. Soc. Am.* **78**, 1958–1965.
- Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.* **102**, 2892–2905.
- Dorman, M. F., Cutting, J. E., and Raphael, L. J. (1975). "Perception of temporal order in vowel sequences with and without formant transitions," *J. Exp. Psychol. Human* **1**, 121–129.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2002). "Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing," *J. Acoust. Soc. Am.* **111**, 2897–2907.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2003). "Recovery from prior stimulation: Masking of speech by interrupted noise for younger and older adults with normal hearing," *J. Acoust. Soc. Am.* **113**, 2084–2094.
- Fogerty, D., Xu, J., and Gibbs, B. E. II (2016). "Modulation masking and glimpsing of natural and vocoded speech during single-talker modulated noise: Effect of the modulation spectrum," *J. Acoust. Soc. Am.* **140**, 1800–1816.
- George, E. L., Festen, J. M., and Houtgast, T. (2006). "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **120**, 2295–2311.
- Gordon-Salant, S., and Fitzgibbons, P. J. (1993). "Temporal factors and speech recognition performance in young and elderly listeners," *J. Speech Lang. Hear. Res.* **36**, 1276–1285.
- Gregan, M. J., Nelson, P. B., and Oxenham, A. J. (2013). "Behavioral measures of cochlear compression and temporal resolution as predictors of speech masking release in hearing-impaired listeners," *J. Acoust. Soc. Am.* **134**, 2895–2912.
- Hopkins, K., and Moore, B. C. (2009). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," *J. Acoust. Soc. Am.* **125**, 442–446.
- Houtgast, T. (1989). "Frequency selectivity in amplitude-modulation detection," *J. Acoust. Soc. Am.* **85**, 1676–1680.
- Howard-Jones, P. A., and Rosen, S. (1993). "The perception of speech in fluctuating noise," *Acta Acust. Acust.* **78**, 258–272.
- Jin, S. H., and Nelson, P. B. (2006). "Speech perception in gated noise: The effects of temporal resolution," *J. Acoust. Soc. Am.* **119**, 3097–3108.
- Jones, M. R., Kidd, G., and Wetzell, R. (1981). "Evidence for rhythmic attention," *J. Exp. Psychol. Human* **7**, 1059–1073.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Kwon, B. J., and Turner, C. W. (2001). "Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference," *J. Acoust. Soc. Am.* **110**, 1130–1140.
- Lentz, J. J., and Marsh, S. L. (2006). "The effect of hearing loss on identification of asynchronous double vowels," *J. Speech Lang. Hear. Res.* **49**, 1354–1367.
- Letowski, T., and Poch, N. E. (1995). "Understanding of time-compressed speech by older adults," *J. Am. Acad. Audiol.* **6**, 433–439.
- Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. (2006). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18866–18869.
- Ma, J., Hu, Y., and Loizou, P. C. (2009). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.* **125**, 3387–3405.
- Miller, G. A., and Licklider, J. C. (1950). "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.* **22**, 167–173.
- Oxenham, A. J., and Kreft, H. A. (2014). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," *Trends Hear.* **18**, 2331216514553783.
- Rhebergen, K. S., and Versfeld, N. J. (2005). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.* **117**, 2181–2192.
- Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.* **120**, 3988–3997.
- Shen, Y., Manzano, N. K., and Richards, V. M. (2015). "Psychometric functions for sentence recognition in sinusoidally amplitude-modulated noises," *J. Acoust. Soc. Am.* **138**, 3613–3624.
- Stine, E. L., Wingfield, A., and Poon, L. W. (1986). "How much and how fast: Rapid processing of spoken language in later adulthood," *Psychol. Aging* **1**, 303–311.
- Stone, M. A., Füllgrabe, C., and Moore, B. C. (2012). "Notionally steady background noise acts primarily as a modulation masker of speech," *J. Acoust. Soc. Am.* **132**, 317–326.
- Stone, M. A., and Moore, B. C. (2014). "On the near non-existence of 'pure' energetic masking release for speech," *J. Acoust. Soc. Am.* **135**, 1967–1977.
- Studebaker, G. A. (1985). "A rationalized arcsine transform," *J. Speech Lang. Hear. Res.* **28**, 455–462.
- Thomas, I. B., Cetti, R. P., and Chase, P. W. (1971). "Effect of silent intervals on the perception of temporal order for vowels," *J. Acoust. Soc. Am.* **49**, 85–85.
- Thomas, I. B., Hill, P. B., Carroll, F. S., and Garcia, B. (1970). "Temporal order in the perception of vowels," *J. Acoust. Soc. Am.* **48**, 1010–1013.
- Vaughan, N. E., and Letowski, T. (1997). "Effects of age, speech rate, and type of test on temporal auditory processing," *J. Speech Lang. Hear. Res.* **40**, 1192–1200.
- Warren, R. M., and Warren, R. P. (1970). "Auditory illusions and confusions," *Sci. Am.* **223**, 30–37.
- Wichmann, F. A., and Hill, N. J. (2001). "The psychometric function: I. Fitting, sampling, and goodness of fit," *Atten. Percept. Psychophys.* **63**(8), 1293–1313.
- Yost, W. A., Sheft, S., and Opie, J. (1989). "Modulation interference in detection and discrimination of amplitude modulation," *J. Acoust. Soc. Am.* **86**(6), 2138–2147.